طرحنامه کرسی علمی – ترویجی

۱. عنوان کرسی:

عنوان فارسی:	**معنا و معیار عاملیت اخلاقی ماشین‌های هوشمند**
عنوان انگلیسی:	The Meaning and Criteria of Moral Agency in Intelligent Machines
عنوان عربی:	معنى ومعايير الوكالة الأخلاقية في الآلات الذكية (ذكاء صناعى)

۳. خلاصه طرحنامه: (حداقل ۱۰۰۰ و حداکثر ۲۵۰۰ کلمه):

الف. موضوع کرسی ترویجی

موضوع اصلی این کرسی با عنوان «معنا و معیار عاملیت اخلاقی ماشین‌های هوشمند» بررسی این پرسش بنیادین است که آیا ماشین‌های هوشمند (Intelligent) و فراهوشمند (Superintelligent) می‌توانند به‌عنوان عاملان اخلاقی (Moral Agents) در نظر گرفته شوند و اساسا معیارهای عاملیت اخلاقی و یا برخورداری از شان اخلاقی چیست؟ این مسئله در تقاطع فلسفه اخلاق فلسفه ذهن، علوم شناختی (Cognitive Science)، و فناوری اطلاعات قرار دارد و به دلیل پیشرفت‌های چشمگیر در هوش مصنوعی (Artificial Intelligence, AI) و تأثیرات گسترده آن بر جامعه، اهمیتی فزاینده یافته است. و در همین راستا، ضمن تمرکز بر تعریف عاملیت اخلاقی، مؤلفه‌های آن، و دیدگاه‌های مختلف درباره امکان یا عدم امکان نسبت دادن این ویژگی به ماشین‌ها، به تحلیل این موضوع می‌پردازد.

عاملیت اخلاقی به توانایی یک موجود برای اتخاذ تصمیم‌های اخلاقی، عمل خودمختار (Autonomous Action)، و پذیرش مسئولیت پیامدهای رفتار و رویه و کنش‌هایش اشاره دارد. در فلسفه، این مفهوم معمولاً به موجوداتی با ویژگی‌هایی مانند خودآگاهی (Self-Consciousness)، اراده آزاد (Free Will)، نیت‌مندی (Intentionality)، و حساسیت اخلاقی (Moral Sensitivity) نسبت داده می‌شود. با این حال، ماشین‌های هوشمند، که بر اساس الگوریتم‌ها و داده‌های از پیش طراحی‌شده عمل می‌کنند، فاقد این ویژگی‌ها به معنای انسانی هستند. این امر پرسش‌هایی را درباره امکان یا چگونگی نسبت دادن عاملیت اخلاقی به آن‌ها مطرح می‌کند. آیا ماشین‌ها می‌توانند تصمیم‌های اخلاقی بگیرند؟ آیا مسئولیت پیامدهای اقداماتشان به خودشان، طراحان، یا کاربرانشان تعلق دارد؟ این پرسش‌ها هسته اصلی مسئله را تشکیل می‌دهند.

ماشین‌های هوشمند، از دستیارهای صوتی مانند Siri تا خودروهای خودران و خودآیین (Autonomous Vehicles)، توانایی‌هایی مانند یادگیری، تصمیم‌گیری مستقل، و تعامل پویا با محیط را نشان می‌دهند. این قابلیت‌ها آن‌ها را از

ابزارهای سنتی متمایز می‌کند، اما آیا این ویژگی‌ها برای عاملیت اخلاقی کافی هستند؟ در این راستا، دو دیدگاه اصلی بررسی می‌شود: نفی عاملیت اخلاقی و پذیرش عاملیت محدود یا عملکردی. دیدگاه نفی، که توسط فیلسوفانی مانند جان سرل (Searle, 1980) و با الهام از ادعا و انگاره ایمانویل کانت حمایت می‌شود، تأکید دارد که ماشین‌ها به دلیل فقدان آگاهی، اراده آزاد، و نیت‌مندی نمی‌توانند عامل اخلاقی باشند. در مقابل، برخی فیلسوفان جدید همچون فلوریدی و سندرز (۲۰۰۴)، با تاکید و تکیه بر مفهوم «اخلاق بدون ذهن» (Mind-less Morality) بر اساس معیارهایی مانند تعامل‌پذیری (Interactivity)، خودمختاری (Autonomy)، و سازگاری (Adaptability) امکان عاملیت محدود را برای ماشین‌ها می‌پذیرد.

این مسئله از منظر عملی نیز اهمیت دارد. تصمیم‌گیری‌های ماشین‌های هوشمند در حوزه‌هایی مانند پزشکی، قضاوت کیفری، و حمل‌ونقل می‌تواند پیامدهای اخلاقی عمیقی داشته باشد. برای مثال، در معضل تراموا / قطار برقی (Trolley Problem)، یک خودروی خودران ممکن است مجبور به انتخاب بین نجات جان عابران یا سرنشینان شود. چنین موقعیت‌هایی نیازمند معیارهای مشخص برای ارزیابی اخلاقی بودن تصمیم‌های ماشین است. علاوه بر این، سوگیری‌های الگوریتمی (Algorithmic Biases) می‌توانند به تبعیض یا آسیب‌های غیرعمدی منجر شوند، همان‌طور که در سیستم‌های پیش‌بینی جرم مشاهده شده است بنابراین، تعریف معیارهای عاملیت اخلاقی نه‌تنها یک چالش نظری، بلکه یک ضرورت عملی برای تضمین عدالت، ایمنی، و اعتماد عمومی به فناوری است.

ابعاد این مسئله شامل بعد شناختی، که به توانایی تحلیل موقعیت‌های اخلاقی اشاره دارد؛ بعد رفتاری، که بر کنش‌های اخلاقی ماشین تمرکز دارد؛ بعد مسئولیت‌پذیری، که به تعیین مسئولیت نتایج می‌پردازد؛ و بعد اجتماعی که به تأثیرات ماشین‌ها بر هنجارهای اجتماعی و فرهنگی مرتبط است. این ابعاد نشان‌دهنده پیچیدگی موضوع و نیاز به رویکردی بین‌رشته‌ای هستند. از منظر فلسفی، بررسی تعارض بین دیدگاه‌های سنتی (مانند کانت و سرل) و دیدگاه‌های مدرن (مانند فلوریدی) از اهمیت ویژه ای برخوردار است و در گرو پاسخ به پرسش‌هایی درباره نقش و جایگاه آگاهی، نیت‌مندی، و حساسیت عاطفی در عاملیت اخلاقی است.

ب. سوابق دانشی: تاریخچه موضوع و سوابق نظری و دانشی و جایگاه موضوع

بحث درباره عاملیت اخلاقی ماشین‌های هوشمند ریشه در فلسفه اخلاق و فلسفه ذهن دارد . افزون بر آغازین گامهایی که کانت (۱۷۸۵) در این زمینه برداشت و عاملیت اخلاقی را به موجودات عقلانی با اراده آزاد و خودمختاری محدود کرد، در قرن بیستم، ظهور هوش مصنوعی با آزمایش فکری تورینگ (۱۹۵۰) پرسش‌هایی درباره توانایی‌های شناختی ماشین‌ها مطرح کرد. سرل (۱۹۸۰) با آزمایش «اتاق چینی» آگاهی و نیت‌مندی ماشینها را از اساس انکار کرد و البته، در مقابل، دنت (۲۰۱۴) امکان عاملیت غیرانسانی را در آینده محتمل دانست و برخی نیز با نگاهی تجدید نظر طلبانه و با تصرف در معنای عاملیت اخلاقی در صدد برآمدند تا ضمن ارایه چارچوبی جدید برای تحلیل این موضوع و توجه به ترجیحات اخلاقی در تصمیم‌گیری‌های ماشین، گونه ای از عاملیت اخلاقی را برای ماشینهای فراهوشمند در نظر بگیرند. این مساله در تقاطع فلسفه، علوم شناختی، و فناوری اطلاعات و هم بسته با چالشهای نظری در باب تعریف آگاهیو همچنین سوگیری الگوریتمی و ... است. پدیده هوش مصنوعی نوپدید و نوپاست و از همین رو در هیچ یک از مقالات منتشر شده به این موضوع مهم پرداخته نشده و تنها در برخی آثار ترجمه شده، اشاراتی به اهمیت و ماهیت این موضوع شده است.

ج. موارد نوآوری:

نوآوری اصلی این کرسی، پرداختن به موضوعی مهم و مبنایی است که در محافل علمی و منشورات دانشگاهی به ندرت بدان اشاره شده است. دیگر وجه نوآورانه این کرسی این است که ناظر به ارائه یک تحلیل جامع و بین‌رشته‌ای از عاملیت اخلاقی ماشین‌های هوشمند که تلفیقی از دیدگاه‌های سنتی و مدرن است و عاملیت اخلاقی را به عنوان یک مفهوم تشکیکی و طیفی در نظر میگیرد و بر اساس تلقی خاص از عاملیت اخلاقی و در حوزه‌های خاص امکان عاملیت اخلاقی ماشینهای هوشمند را نفی و انکار نمیکند. طرح مفهوم «اخلاق بدون ذهن» و طبقه‌بندی عوامل اخلاقی مور (۲۰۰۶) نیز به‌عنوان مبنایی برای گسترش مفهوم عاملیت به موجودات غیرانسانی، از دیگر جنبه‌های نوآورانه این کرسی میتوان به حساب آورد.

د. فرایند:

در گام نخست و با رویکردی تحلیلی مفهوم عاملیت اخلاقی با استناد به منابع فلسفی کلاسیک و و معاصر تعریف و مؤلفه‌های آن (خودآگاهی، خودمختاری، نیت‌مندی، مسئولیت‌پذیری) تبیین میشود. سپس، رویکردها و دیدگاه های اصلی و مشهور در این زمینه به همراه اهم استدلالهای آنها مطرح و بررسی میشود. در پایان و در یک ارزیابی و جمع بندی تحلیلی ضمن برشمردن محدودیت‌های بالفعل و ریشه ای ماشین‌های هوشمند (همچون فقدان حساسیت عاطفی) و امکان عاملیت ماشینهای هوشمند در یک معنای محدود و جدید بررسی و ارزیابی خواهد شد.

هـ. فهرست اهم منابع:

Brożek, B., & Janik, B. (2019). Can artificial intelligences be moral agents? *New Ideas in Psychology*, *54*, 101–106. https://doi.org/10.1016/j.newideapsych.2018.12.002

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Graff, J. (2024). Moral sensitivity and the limits of artificial moral agents. *Ethics and Information Technology*, *26*(1), 13. https://doi.org/10.1007/s10676-024-09755-9

Manna, R., & Nath, R. (2021). The Problem of Moral Agency In Artificial Intelligence. *2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, 1–4. https://doi.org/10.1109/21CW48944.2021.9532549

Misselhorn, C. (2022a). Artificial Moral Agents: Conceptual Issues and Ethical Controversy. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible Artificial Intelligence* (1st ed., pp. 31–49). Cambridge University Press. https://doi.org/10.1017/9781009207898.005

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*(4), 18–21. IEEE Intelligent Systems. https://doi.org/10.1109/MIS.2006.80

Müller, V. C. (2023). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/

Schlosser, M. (2019). Agency. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, *3*(3), 417–457. https://doi.org/10.1017/s0140525x00005756

Sullins, J. P. (2011). When Is a Robot a Moral Agent? In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 151–161). Cambridge University Press. https://doi.org/10.1017/CBO9780511978036.013

Talbert, M. (2024). Moral Responsibility. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2024/entries/moral-responsibility/

خلاصه طرحنامه شامل همه موارد از (الف تا ه) بین ۱۰۰۰ تا ۲۵۰۰ کلمه باشد.